已投稿会议：2024 数智时代知识管理与服务理论、方法、战略学术研讨会

# 基于 BERTopic 模型的国内智慧医疗文献主题挖掘与演化趋势分析

**陈俊冶[1，2] 王康龙[1] 王涟[1] 殷彩明[1] 田园[*2]**

（1. 山西医科大学管理学院　山西 030600；2. 山西医科大学附属肿瘤医院 山西 030600）

**摘要：**［目的/意义］探讨智慧医疗主题相关内容及发展演化趋势，为智慧医疗发展提供参考。［方法/过程］基于知网、万方、维普三大数据库，构建基础词典及停用词表，通过 BERTopic 模型对智慧医疗进行主题特征分析，并建立线性回归模型——以预测未来五年发文量。［结果/结论］智慧医疗在未来五年发文数量持上升趋势，并涵盖智能化、数字化、个性化三个方向。研究有助于医院加强基础设备的智能化建设，健全相关法律法规及监管机制，国家加强支持力度及数字化设备的推广以及该领域的人才培养，推进智慧医疗的快速发展。

**关键词：**智慧医疗　BERTopic　线性回归方程　主题特征

"智慧医疗"指通过利用可穿戴电子设备、物联网和移动网等设备来连接人员、资源和组织的卫生系统平台[1]，并借助机器学习、深度学习等算法技术[2]对疾病进行智能决策的过程[3]，其建设载体为智慧医院[4]。2017 年底，国家卫健委与中医局联合发布《改善医疗服务计划（2018 至 2020 年）通知》，强调利用"互联网＋"革新医疗模式，构建智慧型医院；随后，2020 年 4 月，国家卫健委发布《关于进一步完善预约诊疗制度，推进智慧型医院建设的通知》，明确将实现医疗、服务及管理一体化的智慧型医院作为未来建设方向；2021 年，国办又出台《公立医院高质量发展意见》，强调以信息技术驱动医院高质量发展[5]。基于"健康中国 2030"规划，医疗行业正在经历从传统医疗到数字医疗、信息医疗，再通过数字医疗向智慧医疗转变[6]。近年来，智慧医疗已然成为我国医疗行业发展的工作重心，这对于推动健康中国建设，促进人民更智能、更便捷就医起着举足轻重的作用。本文通过使用 BERTopic 模型对我国智慧医疗进行主题挖掘以及发展演化分析，旨在为相关学者通过借鉴与参考，推动该主题的深入探讨。

## 一、 资料与方法

### 1.1 数据来源

本研究选取知网（CNKI）、万方(WanFang)和维普（VIP）三大数据库为数

通讯作者：田园，山西医科大学附属肿瘤医院信息科，主任技师。

据源，发表时间限定为“2024 年 1 月 31 日”，选取检索词“智慧医疗”，进行精确检索，共检索到文献 8150 篇。

## 1.2 数据预处理

本研究数据处理有以下步骤：1.通过 NoteExpress 软件将三个数据库导出的数据进行去重处理，并将题录通过 excel 格式导出。2.将不完整数据，如无发表时间、无摘要或摘要不完整，主题不相关等数据进行剔除。3.本研究通过使用 jieba 库对中文文本进行分词，其中包括了建立词典的分词、停用词去除和基本的过滤技术进行数据预处理。

## 1.3 研究方法

本文首先使用通过使用 pandas 库提取 Excel 文件中文献发表数据，并将其存储在 DataFrame 中，通过 sklearn 库中的 LinearRegression 线性回归模型对年份和发表数量及进行拟合，预测该主题后续的发表数量。

其次，本文采用基于 BERT 主题建模算法[7]BERTopic 进行主题识别和发展演化分析。其中，BERTopic 模型可以理解为一个模块的集成管道，主要包括文本嵌入、数据降维、聚类和主题表示[8]，见图 1。该模型首先通过 paraphrase-multilingual-MiniLM-L12-v2 进行文档嵌入、UMAP 算法降低嵌入的维数、HDBSCAN 算法创建语义相似文档的聚类以及 c-TF-IDF 算法进行候选关键词的提取，然后对该模型进行训练，通过 fit_transform 对输入文本向量化、Topic_model 模型提取主题 Topics，并且计算主题文档概率 probabilities，然后使用 berTopic 模型进行可视化分析，最后使用 DTM 模型对获取不同主题随时间推移的可视化模型。

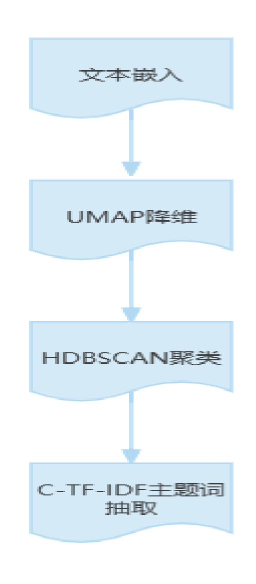



图 1 BERTopic 主题建模

# 二、 数据分析

## 2.1 文献发文情况

本研究通过使用 python 语言与线性回归模型可视化智慧医疗发文情况以及预测未来五年的发文情况。如图 2，智慧医疗相关文献于 2009 年初次被发表，发文数量呈逐年波动上升趋势，并于 2021 达到顶峰，于 2023 年出现下滑（因 2024 年文献只有一月份，不做参考）。并针对该主题通过使用 sklearn 库中的 LinearRgression 类，建立线性回归模型，预测该主题文献在未来五年的发表数量将会持上升趋势。综上，智慧医疗领域已经引起学术界的广泛关注。

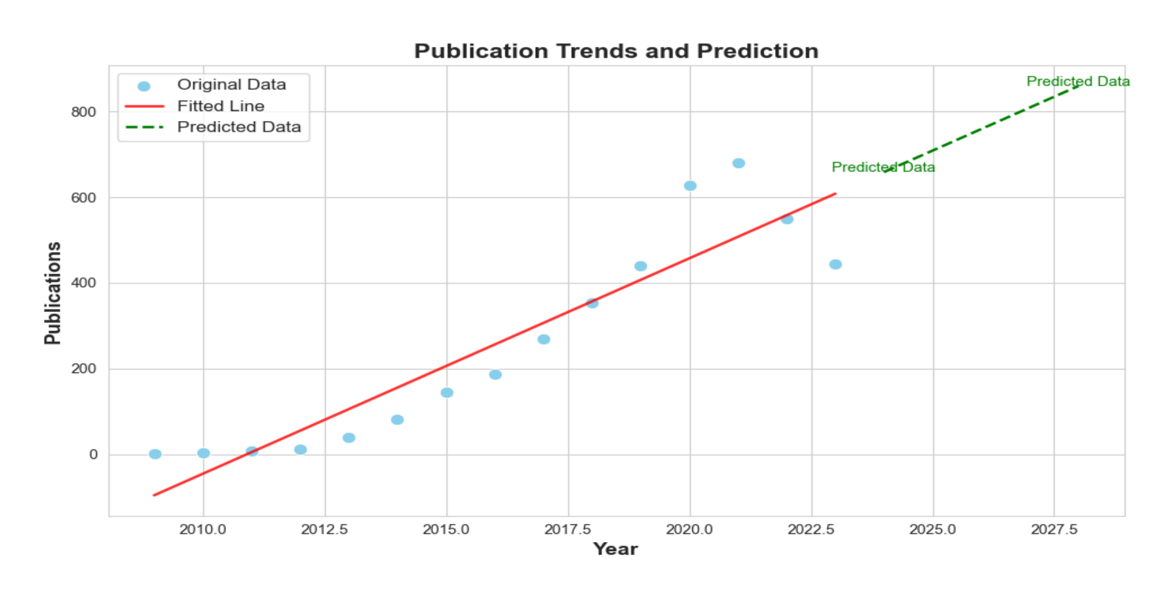



图 2 智慧医疗领域发文量及未来五年发文预测

## 2.2 热点话题识别

本研究通过运用 BERTopic 模型获取智慧医疗相关的 8 个显著性较高的主题，及这些主题中占比前五的关键词，便于后续的分析。如图 3，通过这些关键词可以将这八个显著性较高的主题分别概括为数字化医疗生态系统、智慧老年健康管理、数字化转型、综合服务医疗网络、医疗服务体验、智能化医疗技术应用、数字化健康管理、智能医疗云服务与管理。其中这 8 个主题分别代表智慧医疗的四个方向.分别是：基础设施建设和数字化转型、健康管理与网络服务、医疗技术与应用，医疗体验。由此可知，当前智慧医疗当前朝着数字化、人性化、智能化，多方面多维度发展。




图 3 智慧医疗主题词分布可视化

## 2.3 层次聚类

使用 BERTopic 模型进行层次聚类，可以更好了解每个主题之间的关系，并提高数据的可解释性。如图 4，其中 Topic24、Topic10 与 Topic8 体现智慧医疗在技术方面的融合与创新，Topic18 与 Topic2 体现智慧医疗的基础建设以及对经济的影响，Topic27 与 Topic22 体现银行与区块链技术在医疗中的应用与创新，Topic21、Topic5 与 Topic16 体现智慧医疗提高使用一些新的技术提高疾病诊治的准确率，Topic16、Topic19、Topic15 与 Topic20 则体现出智能医疗背景下，提高医院信息系统与智能设备的结合，提高医疗质量，Topic30 与 Topic26 体现医疗信息化建设中档案管理的创新，Topic28、Topic17 与 Topic13 体现不同不同机构对医院信息化建设的重视以及所作出的贡献，Topic29 则体现出智慧医疗的快速发展以及带来的便捷，Topic11 与 Topic6 则体现出智慧医疗在促进健康中国发

挥着举足轻重的作用，Topic25、Topic14、Topic4、Topic3、Topic9、Topic0 与 Topic7，Topic12 体现智慧医疗的智能化、多元化、人性化、数字化，Topic1 与 Topic23 体现智慧医疗关注慢性疾病、老年人及看病不易的医疗问题。

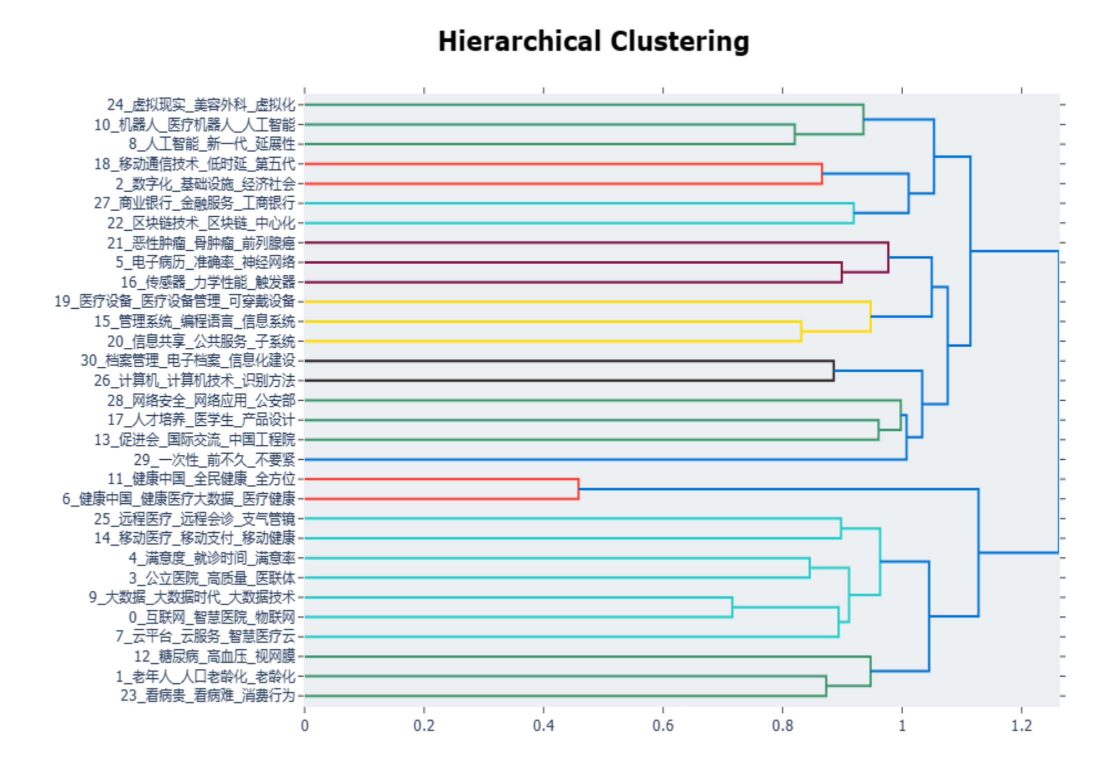



图 4 层次聚类图

## 2.4 主题相似度热力图

本研究使用 HDBSCAN 算法生成相似度热力图。主题之间的相似度或者相关性是通过颜色的深浅表现[9]。如图 5，Topic0 与每一个主题之间相关性较高，可见智慧医疗是通过互联网、物联网、智慧医院等关键词为载体进行的。Topic6 与 Topic11 呈现较高的相关性，可见医疗机构通过使用信息技术、大数据以及智能化医疗等手段促进全社会的健康管理，提高医疗的服务质量。

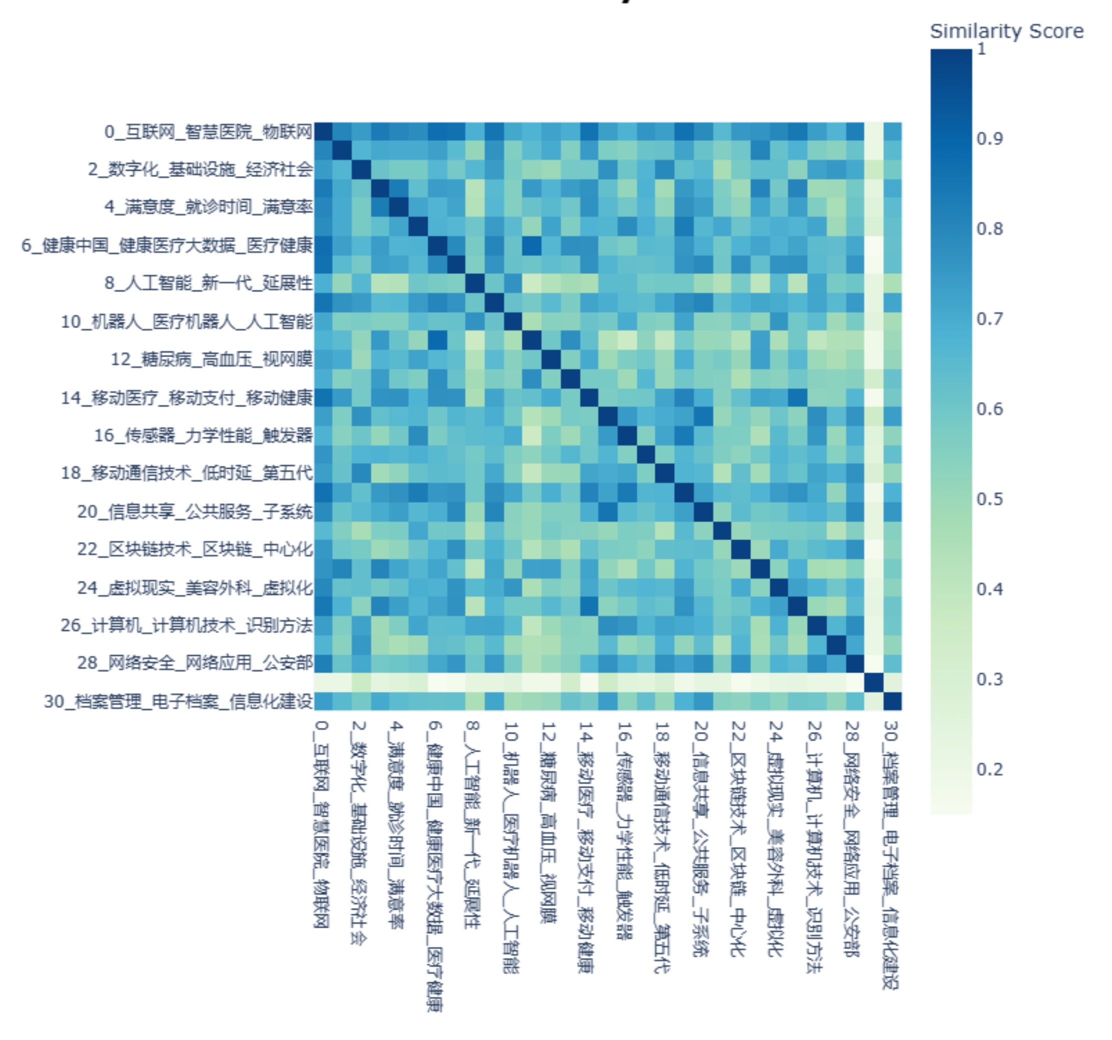



图 5 主题相似热力图

## 2.5 研究主题发展演化趋势

本研究使用 DTM 对各主题进行发展演化趋势分析，使用折线图以及不同颜色呈现，便于对不同主题进行比较及理解[10]。如图 6，Topic0 一直持波动递增趋势，于 2019 年达到顶峰，后呈波动下降趋势。Topic1、Topic9 于 2013 年之后出现呈缓慢递增发展，其中 Topic1 在 2017 年以及 2021 年达顶峰后呈下降趋势发展。其余主题与 2010 年之后陆续出现，并呈缓慢发展趋势。

由图可知，2019 年之后，大部分主题呈下降趋势亦或是发展缓慢，这有可能是因为疫情的原因，致使研究人员重心转移，或因为技术相对成熟，需攻克更为复杂的技术，周期变长，抑或是因为支持力度减弱。Topic1 是因为中国的

人口老龄化日益突现，引起相关学者关注。

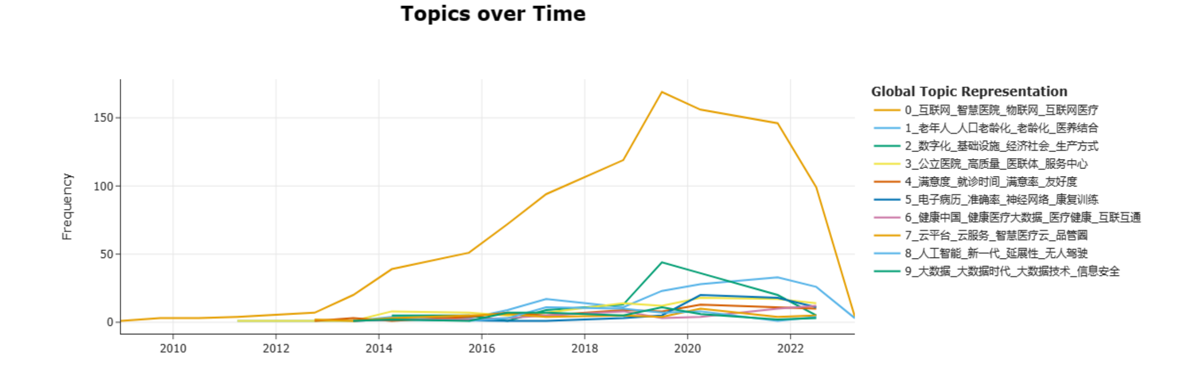



图 6 研究主题发展演化趋势

## 三、 讨论

本研究通过针对主题关键词、层次聚类、以及主题相似度热力图可知，智慧医疗包含了智能化、个性化以及数字化三个方向。因此，本研究从这三个方向展开讨论。

### 3.1 智慧医疗智能化

智慧医疗通过运用云计算、5G、人工智能、物联网等新技术，可以实现对初次入院患者进行多维度患者画像[11]；开展多方面的远程医疗[12]，如远程医疗、远程手术、远程会诊等，并未平台患者就提通过了便捷[13];智慧医疗的智能化使得医疗服务水平大幅提升，医疗失误和过度医疗等情况有效减少，为实现了医疗资源的精细化管理奠定坚实基础。同时，随着智慧医疗的智能水平的提升，医疗资源的利用率得以提升，医院人力物力成本不断减少，医疗资源的分配和利用更加合理，医患关系得以改善。

### 3.2 智慧医疗个性化

个体间的差异性导致在治疗疾病个体的过程中须根据其个体独特的生化、生理、环境暴露和行为特征[14]进行诊断、预防和治疗与健康相关的疾病 [15]并对其制定有针对性的治疗和预防计划[16]。个性化医疗通过利用人工智能和大数据等技术使患者获得更加高效、安全、精准的医疗服务的方式。如 Dignio 平台通过收集患者身体指标并对其分配相适应的药物,Medicio 可以为患者提供处方的

多种药物的个性化工具包[17],通过平台个性化医疗服务对患者的病史、症状、生理指标等数据进行深度分析和挖掘，为患者提供个性化的医疗服务，提高医疗精准度。

### 3.3 智慧医疗数字化

智慧医疗的数字化包括远程医疗、医疗大数据、深度学习、电子病历等方面[18]。电子病历是数字化医疗的核心，它可以帮助医生更好地了解患者的疾病情况，制定治疗方案，也能够实现医疗资源的共享和患者数据的安全管理。此外，远程医疗也是智慧医疗数字化的一个重要方面，它可以通过网络实现医生和患者的远程交流和诊断，从而提高医疗服务的效率和质量。而医疗大数据的应用则可以帮助医院和医疗机构更好地了解患者的疾病情况，优化医疗服务流程，提高服务质量和效率。医疗大数据的应用也为医疗行业的未来发展提供了更多的可能性，例如，通过对大量患者数据的分析和挖掘，可以发现新的疾病治疗方法和药物，为患者的治疗带来新的希望。

## 四、 总结与建议

### 4.1 研究总结

本研究通过 berTopic 模型对智慧医疗主题进行挖掘及可视化分析，发现智慧医疗在未来五年的相关文献发文量持上升趋势，及涵盖了智能化、个性化、数字化三个方向，并以信息技术、人工智能、数字化等技术为支撑提高医疗服务水平。

智慧医疗的相关主题呈多元化，但 2022 年之后都出现不同程度的下降趋势，这可能与技术的挑战、数据的隐私、该领域的相关法律法规和监管环境以及成本和可持续等原因有关。

### 4.2 研究建议

首先，在基础设施建设方面，应积极加强智能化基础设施的建设，推广相关设备与及系统的应用，积极加强智能设备与系统的引进与研发工作，提高医疗技术水平和医疗质量。还需对医护人员进行智能技术培训，提高他们对智能化设备和系统的认识和理解，从而最大化发挥智能化设备和系统的作用，以提

高诊断准确率和治疗效果。

其次，应建立健全相关的法律法规及监管制度。不同医疗机构之间的数据互通和共享，确保数据的安全性。建立完善的数据清洗和验证机制，为后续的数据分析提供强有力的保障。

此外，国家应加强相关政策保障以及财政支持力度鼓励医疗机构使用数字化工具和技术，如电子病历、远程医疗、移动医疗、可穿戴设备等，提高医疗服务的效率和质量。

最后，应健全智慧医疗服务评价体系，通过线上平台亦或者线下问卷、电话访问等方式，获取患者就医体验，医疗机构应及时总结问题及解决措施加以改正。并加强相关人才培养，加快智慧医疗发展。

## 参考文献

[1] Ziwei H ,Dongni Z ,Man Z , et al.The applications of internet of things in smart healthcare sectors: a bibliometric and deep study[J].Heliyon,2024,10(3):e25392-.

[2]蔡大顺.智慧医疗民事责任的内容革新与赔偿优化[J].江汉论坛,2023(12):121-126.

[3]唐旭,郭宇飞,陈曦等.智慧医疗环境下老年慢性病病人技术焦虑现状及影响因素[J].护理研究,2023,37(21):3925-3930.

[4]曾建丽,王华欢,马瑞晨等.智慧医疗服务满意度评价及影响因素研究[J].中国医院,2022,26(06):42-44.DOI:10.19660/j.issn.1671-0592.2022.6.13.

[5].智慧医疗如果更智慧[J].中国卫生,2023(11):20-21.

[6]周罗晶,邵旸,张瑞等.智慧医疗场景下人工智能应用伦理问题与治理路径探讨[J].中国医院,2024,28(02):38-41.DOI:10.19660/j.issn.1671-0592.2024.2.10.

[7]徐汉青,滕广青.机遇与挑战：基于 BERTopic 的 AI 环境下图书馆主题文本挖掘[J/OL].情报科学:1-20[2024-02-25].http://kns.cnki.net/kcms/detail/22.1264.G2.20231103.1007.014.html.

[8]Tang, Z., Pan, X., & Gu, Z. (2024). Analyzing public demands on China's online government inquiry platform: A BERTopic-Based Topic modeling study. PloS one, 19(2), e0296855. https://doi.org/10.1371/journal.pone.0296855

[9]王辉,王晓玉,顾东晓等.在线健康社区重大慢病患者负面评论倾向的关键影响

因素分析[J/OL].情报科学,1-19[2024-02-24].http://kns.cnki.net/kcms/detail/22.1264.g2.20240129.0942.010.html.

[10]高春玲,姜莉媛,董天宇.基于 BERTopic 模型的老年人健康信息需求主题演化研究——以新浪微博平台为例[J/OL].情报科学:1-16[2024-02-25].http://kns.cnki.net/kcms/detail/22.1264.G2.20240128.1743.004.html.

[11]计虹.人工智能助力智慧医院高效诊疗[J].中国卫生,2023(11):22-23.DOI:10.15973/j.cnki.cn11-3708/d.2023.11.013.

[12]周巍.基于 5G 技术的智慧医疗应用仿真技术探析[J].电脑知识与技术,2023,19(30):104-107.DOI:10.14004/j.cnki.ckt.2023.1615.

[13]张永超. 智慧医疗触“云”可及 数字健康链接未来[N]. 中国城市报,2024-02-05(014).

[14]Goetz, L. H., & Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. Fertility and sterility, 109(6), 952–963. https://doi.org/10.1016/j.fertnstert.2018.05.006.

[15]Salari, P., & Larijani, B. (2017). Ethical Issues Surrounding Personalized Medicine: A Literature Review. Acta medica Iranica, 55(3), 209–217.

[16]Pelter, M. N., & Druz, R. S. (2024). Precision medicine: Hype or hope?. Trends in cardiovascular medicine, 34(2), 120–125. https://doi.org/10.1016/j.tcm.2022.11.001

[17]Kataria, S., & Ravindran, V. (2022). Musculoskeletal care - at the confluence of data science, sensors, engineering, and computation. BMC musculoskeletal disorders, 23(1), 169.https://doi.org/10.1186/s12891-022-05126-x.

[18]Yeung, A. W. K., Torkamani, A., Butte, A. J., Glicksberg, B. S., Schuller, B., Rodriguez, B., Ting, D. S. W., Bates, D., Schaden, E., Peng, H., Willschke, H., van der Laak, J., Car, J., Rahimi, K., Celi, L. A., Banach, M., Kletecka-Pulker, M., Kimberger, O., Eils, R., Islam, S. M. S., … Atanasov, A. G. (2023). The promise of digital healthcare technologies. Frontiers in public health, 11, 1196596. https://doi.org/10.3389/fpubh.2023.1196596